# Deep learning for Parkinson's Disease diagnosis based on handwriting examination

Internship Report


**Author**: Paul Lerner
**Supervisor**: Laurence Likforman-Sulem

Département Image, Données, Signal
Laboratoire Traitement et Communication de l'Information
Télécom Paris

August 30, 2019

# Acknowledgments

I would like to thank Laurence for her counsel throughout my internship, it was a pleasure working with you. I thank also the rest of the people from *Télécom*, especially Tanvi, for the friendly atmosphere.

I also want to thank my brother Ivan for his advice and my girldfriend Louise for her support.

This work was funded by the Data Science & Artificial Intelligence for Digitalized Industry & Services[1] chair.

[1] https://www.telecom-paris.fr/fr/recherche/recherche-partenariale/les-chaires-de-recherche/data-science-and-artificial-intelligence-for-digitalized-industry-and-services

# Abstract

Parkinson's Disease (PD) is the second most common neurodegenerative disease. Its diagnosis is considered to be difficult as there is a large number of symptoms which are shared among other diseases. Several of its symptoms are observable through handwriting analysis at an early stage of the disease.

A few authors have already used machine learning techniques in order to classify PD patients (PDs) and healthy controls based on hand-crafted features extracted from handwriting examinations. The goal being to provide for a Clinical Decision Support System (CDSS). However it is very difficult to decide which feature to extract and some inconsistent results have been obtained depending on the handwriting task. Therefore we decided to focus on deep learning models which are able to learn features from the raw data.

We review the literature related to our work, namely CDSS for PD. We first describe works based on handwriting data as ours. Those provide evidence that handwriting impairments are early markers for PD. Moreover, some provide encouraging results for the monitoring of PD. We then go over other examinations which have been used to built CDSS. Vocal analysis allowed for a great sensitivity.

We propose to use uni-dimensional convolutional neural networks to learn PD-discriminative features from online handwriting data. Our model achieves competitive results with the literature, however, it lacks for explainability.

We go over several limitations that we share with the rest of the literature : medicated PDs, universality of the examination, differential diagnosis. The perspectives of our work rely on transfer learning and explainable models, along ways to address these different limitations.


*The work presented in this report was done in the context of my internship at the* Laboratoire Traitement et Communication de l'Information *of* Télécom Paris*, thus fulfilling my studies at* Université Paris Descartes. *My work was supervised by Laurence Likforman-Sulem, associate professor and HDR at* Télécom Paris.


***Keywords*** : Parkinson's Disease · Deep learning · end-to-end learning · Online handwriting · Handwriting features · Clinical decision support system

# Contents

# 1 Introduction

> *"Hitherto the patient will have experienced but little inconvenience; and befriended by the strong influence of habitual endurance, would perhaps seldom think of his being the subject of disease, except when reminded of it by the unsteadiness of his hand, whilst writing or employing himself in any nicer kind of manipulation. But as the disease proceeds, similar employments are accomplished with considerable difficulty, the hand failing to answer with exactness to the dictates of the will."*
>
> — Parkinson (1817)

Parkinson's Disease (PD) is the second most common neurodegenerative disease (Tanner et al., 1999). Its diagnosis is considered to be difficult as there is a large number of symptoms which are shared among other diseases (see, e.g. Hughes et al. (2002)). However, the diagnosis is usually assessed by a clinician after a – relatively subjective – physical examination as SPECT and CT scans are costly, invasive, and usually effective when the disease has already progressed to a mature stage (Moetesum et al., 2019).

As the disease is described in Parkinson (1817), writing deficits precede walking deficits (see the above quote). Since then, handwriting has been used to assess PD through tasks like Archimedean spiral or simple subjective rating like in the Unified Parkinson's Disease Rating Scale (UPDRS, Goetz et al.). Symptoms such as *tremor* or *micrographia* (i.e. diminished letter size) would be visible through a traditional paper-and-pencil examination (offline handwriting). Moreover, McLennan et al. (1972) found that *micrographia* may antedate additional motor signs of PD by three to four years. Furthermore, researchers now argue that PD *dysgraphia* is larger than *micrographia* and that it is observable through *kinematic* analysis : Letanneux et al. (2014) report that writing velocity and smoothness/fluency abnormalities are more frequent (above 75%) than diminished letter size (between 30%-50%). In addition, the use of smart pens and digital tablets has allowed to collect online handwriting data, thus, to extract a large number of features, on which to perform statistical analysis and machine learning. The goal for the researchers is to provide for a Clinical Decision Support System (CDSS) which would confirm or question the neurologist' diagnosis based on an inexpensive and non-invasive handwriting examination.

Several authors have already used machine learning techniques in order to classify PD patients (PDs) and Healthy Controls (HCs) based on hand-crafted features extracted from the data. However PD's handwriting impairments result from numerous symptoms : *akinesia*, *bradykinesia*, *rigidity*, *tremor* and *micrographia*/*dysgraphia* (Phillips et al., 1991; Letanneux et al., 2014). *"Since these symptoms are relatively independent of each other (Zetusky et al., 1985), they probably involve different mechanisms (for example, Parkinsonian tremor is associated with cholinergic rather than*

*dopaminergic neurotransmitter systems; (Stahl, 1986)).”* — Phillips et al.. Moreover, these symptoms are not consistent among all PDs, e.g. *micrographia* is only present in approximately 30%-50% of PDs (Letanneux et al., 2014). Because of this, it is very difficult to decide which feature to extract and some inconsistent results have been obtained depending on the handwriting task (Drotár et al., 2016). Therefore we decided to focus on deep learning models which are able to learn features from the raw data. Deep learning has brought breakthroughs in automatic speech recognition and computer vision (LeCun et al., 2015), its application to healthcare is an active research field (Miotto et al., 2017).

The work presented in this report was done in the context of my internship at the *Laboratoire Traitement et Communication de l'Information* (LTCI) of *Télécom Paris*, thus fulfilling my studies at *Université Paris Descartes*. My work was supervised by Laurence Likforman-Sulem, associate professor and HDR at *Télécom Paris*.

*Télécom Paris* (formerly known as *Télécom ParisTech*) is one of the top French engineering school, founded in 1878[2]. As for it, the LTCI was funded in 1982 and comprises 130 researchers and as many PhD students[3]. Both are located in Paris but will soon move to *Plateau de Saclay*.

Laurence Likforman-Sulem is part of the *Image, Données, Signal* (IDS) department among LTCI. The research interests of IDS include[4] :

- Signal and images analysis and processing (audio, video, multimedia, satellite images, biomedical images, ...)
- Analysis and development of statistical processing methods and algorithms for machine learning, optimisation and data analytics.

Laurence received her PhD from *Télécom Paris* in 1989 and her HDR from *Université Pierre & Marie Curie*. Her research interests include[5]:

- document analysis dedicated to handwritten and historical documents
- document image understanding
- character recognition

She has recently worked on emotional state recognition from online handwriting and drawing data (Likforman-Sulem et al., 2017). Also, she supervises the PhD of Catherine Taleb which aims at using multimodal data for a PD CDSS (Taleb et al., 2018).

[2]https://www.telecom-paris.fr/fr/lecole/telecom-paris-en-bref/histoire

[3]https://www.telecom-paris.fr/en/research/laboratories/information-processing-and-communication-laboratory-ltci

[4]https://www.tsi.telecom-paristech.fr/en/

[5]https://perso.telecom-paristech.fr/lauli/engbiogr.htm

This report is organized in 6 sections, including this introduction. We will first introduce some background materials necessary to understand the rest of this report. Then, we will review the literature related to our work – namely CDSS for PD – before justifying the choice of the database used in the penultimate section : *Machine learning*. The latter describes the proposed model architectures which are then evaluated and compared to the related works. Finally, we will conclude by summarizing our different contributions, discussing the limitations and perspectives of our work.

# 2 Background materials

This section presents several technical domains and techniques necessary to understand the rest of this report.

## 2.1 PD symptoms

In this section we define a few PD symptoms which are *a priori* observable through handwriting analysis.

- *Micrographia* : abnormal reduction in writing size (Letanneux et al., 2014)
- *Dysgraphia* : handwriting impairment in PDs (Letanneux et al., 2014).
- *Tremor* : involuntary to and fro movements that can be visualized by irregular formations of characters and drawings (Moetesum et al., 2019).
- *Rigidity* : stiff or inflexible muscles[6].
- *Dyskinesia* : uncontrollable, often jerky movements that a person does not intend to make. These movements can affect the arms, legs, head or whole body[6].
- *Bradykinesia* : slowness of movement which causes PDs to complete a graphomotor task in more time than usually required (Moetesum et al., 2019).
- *Akinesia* : loss of physical movement.

We will use these terms throughout the report, especially in sections 3.1 and 4. The next sections are more technical as they relate to statistics and machine learning.

## 2.2 Statistics

Student's t-test is often used to determine if the means of two sets of data are significantly different from each other, we use it in section 4.2 to analyze databases. It can only be applied to data which :

1. has an equal variances in both its sets (Bartlett's test).
2. is drawn from a normal distribution (Shapiro-Wilk's test).

## 2.3 Machine learning

In this section we present a few supervised-learning models which are used in the rest of this report.

[6]https://www.parkinsonsvic.org.au/

### 2.3.1 Linear Discriminant Analysis (LDA)

In section 4.2, we use Linear Discriminant Analysis (LDA) in order to classify PD and HC based on specific features.

LDA models the distribution of the features $X$ separately in each of the response classes (i.e. given $Y$, in our case PD or HC), and then use Bayes' theorem to flip these around into estimates for $P(Y = k|X = x)$.

### 2.3.2 Neural networks

A neural network is a collection of units called neurons. Neurons are connected to each other by a *weighted* connection. Each neuron has an activation value which is equal to $f(\sum_{i=1}^{n} x_i w_i)$. Where $f$ is its activation function, $x_i$ is another neuron connected through a $w_i$-weighted connection. Neural networks are able to learn by adjusting those *weights*. This is usually done via Stochastic Gradient Descent (SGD) which computes the derivative of a given *loss* function w.r.t. the *weights* of the network.

Neural networks are usually organized in several layers. Neurons among a layer are not connected to each other. A Fully Connected (FC) layer is a layer of neurons which are all connected to the previous layers' neurons.

In section 5 we use Convolutional Neural Networks (CNNs) to learn PD-discriminative features from online handwriting data. CNNs were first applied to optical character recognition by LeCun et al. (1998), see figure 1. The three main architectural ideas are :

1. local receptive fields : each neuron in a layer is connected to a small neighbourhood of neurons (the *kernel*) in the previous layer. This was inspired by the cat's visual system (Hubel and Wiesel, 1962).

2. shared weights : each feature map is designed to extract one feature from the previous layer as its weights are constrained to be identical thus performing the same operation on different parts of the image.

3. spatial or temporal sub-sampling : performs a local *max-pooling*, reducing the resolution of the feature map, thus reducing the sensitivity of the output to shifts and distortions.

CNNs are now state-of-the-art in most computer vision tasks such as optical character recognition, face recognition and image segmentation (LeCun et al., 2015). Although originally thought for and mainly applied on 2D images, CNNs are also applicable to time-series in their uni-dimensional form :

- Kim (2014) apply it on sentence classification (mostly sentiment analysis)

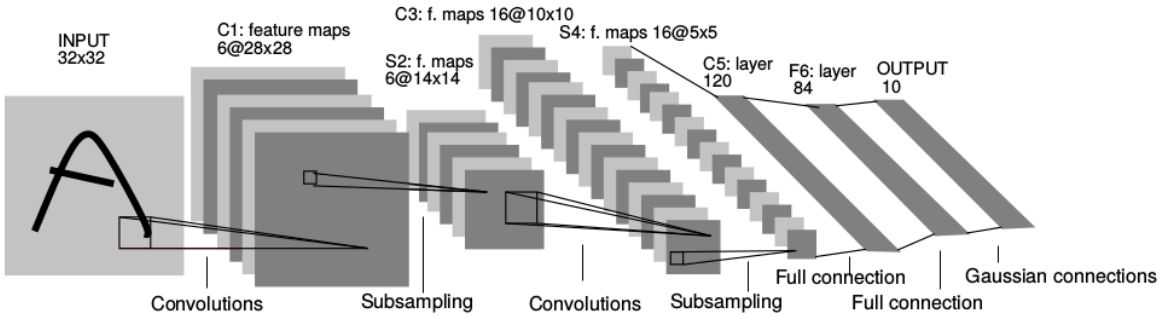


Figure 1: From LeCun et al. (1998): Architecture of LeNet-5, a CNN designed for Optical Character Recognition, *"each plane is a feature map, i.e. a set of units whose weights are constrained to be identical"*.

- Bai et al. (2018) apply it on several sequence modeling tasks (e.g. language modelling, music modelling) and outperform Recurrent Neural Networks on most of those.
- Oh et al. (2018) apply it on Electroencephalography (EEG) signals to diagnose PD, see section 3.2.

These last works have inspired us to use CNNs for PD diagnosis, see section 5.1.

### 2.3.3 Majority voting

Majority voting is an ensemble method which allows to combine the predictions of several predictive models : the final (elected) prediction is the one which has been predicted by the majority of the models. This technique is notably used by Pereira et al. (2018); Moetesum et al. (2019) and ourselves (see sections 3.1 and 5).

### 2.3.4 Dataset split

In machine learning, datasets are usually split in two :

1. a *training* set which will be used to train the model (using, e.g. SGD).
2. a *test* set which will be used to evaluate the model at the end of the training (by, e.g., computing the classification accuracy).

The goal of the model is to *fit* the data of the training set. When the model's performance is significantly higher on the training than on the test set, we say it is *overfitting*. It means that the model *learns noise* from the training set instead of relevant features.

In order to provide statistically stronger results – especially on small datasets – a common practice is $k$-fold Cross-Validation ($k$-CV). $k$-CV consists in splitting the

dataset into $k$ folds, train the model over the $k$ first folds and test it on the last one. The operation is repeated $k$ times until every fold has been used to evaluate the model. The results are then averaged over the $k$ folds.

10-CV is notably used by Drotár et al. (2016); Moetesum et al. (2019); Mucha et al. (2018), and, hence, ourselves (see sections 3.1 and 5.2.2).

## 2.4 Evaluation metrics

In order to evaluate the performance of predictive models in medicine, one should assess for its sensitivity and specificity, defined below (Bossuyt et al., 2015). Therefore we use those in section 3 to compare the results of authors working on different databases. However, in section 5.2, since we work on a perfectly balanced dataset, we assess the performance of our model using simply its classification accuracy ($Acc$), which, in this case, is the average between sensitivity and specificity.

Sensitivity ($Se$, aka *recall* aka *true positive rate*) refers to the test's ability to correctly detect ill patients who do have the condition.

Specificity ($Sp$, aka *true negative rate*) relates to the test's ability to correctly reject healthy patients without a condition (Altman and Bland, 1994).

If we define :

- True positive ($TP$) as sick people correctly identified as sick
- False positive ($FP$) as healthy people incorrectly identified as sick
- True negative ($TN$) as healthy people correctly identified as healthy
- False negative ($FN$) sick people incorrectly identified as healthy

then :

$$Se = \frac{TP}{TP + FN} \; ; \; Sp = \frac{TN}{TN + FP} \text{ and } Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{1}$$

In the next section we will make use of these evaluation metrics to compare the different performances of existing CDSS.

# 3 Related works

This section reviews the literature on topics related to our work, namely CDSS for PD. We will first describe works based on handwriting data as ours, then, go over other examinations which have been used to built CDSS.

## 3.1 CDSS for PD via handwriting examination

This section reviews works related to CDSS for PD via handwriting examination. We will start by reviewing the different handwriting tasks which are given as an examination, then, we will describe two widely used and publicly available databases before reporting the different features and machine learning models used to classify these data. This last part is divided into two : we first review the different hand-crafted features along with their model then we describe the methods used in deep learning approaches.

### 3.1.1 Handwriting tasks for PD's assessment

Table 1: Summary of the different handwriting tasks. If several authors use the same database, only one (who collected the data) is cited. *Subjects had to write over a template.

| **Task** | **Authors** |
|---|---|
| Static Archimedean spiral | Drotár et al. (2016), Stanley et al. (2010), Saunders-Pullman et al. (2008), San Luciano et al. (2016), Pereira et al. (2018)*, Graça et al. (2014)*, Isenkul et al. (2014)*, Smits et al. (2014)* |
| Dot-template Archimedean spiral | Zham et al. (2017)* |
| Dynamic Archimedean spiral | Isenkul et al. (2014)* |
| Repetitive-cursive letter (e.g. letter *l*) | Drotár et al. (2016), Taleb et al. (2017), Bidet-Ildei et al. (2011), Smits et al. (2014)* |
| Miscellaneous words and sentences | Drotár et al. (2016), Rosenblum et al. (2013b), Taleb et al. (2017) |

The design of handwriting tasks for the study of PD is delicate as several factor may influence the symptoms of the patients. Berardelli et al. (2001) show that paying attention to movement is beneficial to *bradykinesia*. In the same way, Wu et al. (2015) found that asking patients to pay attention to the letter size improves *micrographia*. Moreover, Letanneux et al. (2014) recommend that subjects do not write a long text because it implies either copying it or writing under dictation, and in both cases, cognitive processes are required.

Drotár et al. collected the *PaHaW* database (described in the next section) which we have decided to work on throughout my internship (see section 4). They designed

their tasks based on the works of Stanley et al. (2010) – who suggest that spiral analysis may be more sensitive in detecting early disease than the usually assessed UPDRS – and Smits et al. (2014) who were able to measure *bradykinesia*, *tremor* and *micrographia* thanks to a repetitive-cursive letter task. A summary of the different tasks used in the literature is available in table 1.

From table 1, one can observe two different divergences from the literature :

1. The use of a template for the handwriting task.

2. The design of the said template for the Archimedean spiral.

Indeed, Zham et al. (2017) argue that spiral drawing without template leads to significant inter-participant variability. If this might not be a problem for the works of Drotár et al. where each subject only performs one task, it might for the works of, e.g. San Luciano et al..

Moreover, Zham et al. also argue that a dynamic template that appears and disappears at certain time intervals (as in Isenkul et al. (2014)) is unsuitable for elderly patients. Therefore they opt for a light template consisting of a dotted-line instead of a full-line. Isenkul et al. do not report that the dynamic template might be unsuitable for the elderly and come to the conclusion that both static and dynamic tasks can be used together in order to assess for PD.

Despite these two divergences, one can see from table 1 that the tasks used by Drotár et al. are widely established, especially the spiral task. We will further describe the *PaHaW* database in the next section.

### 3.1.2 Databases description

In this section we present two widely used and publicly available databases that we study in section 4 : *PaHaW* and *NewHandPD*. See also appendix A for technical specifications about the databases.

The *PaHaW* database comprises 75 subjects (37 PDs and 38 HCs), although 1 PD and 2 HCs were discarded from this study because they did not perform the spiral task. Therefore in this study we use a perfectly balanced dataset of 36 HCs and 36 PDs (as did Moetesum et al.; Impedovo et al.). The examination consists of seven handwriting tasks (see Figure 2). From the first to the third task, participants wrote cursive letters or bi/tri-grams of letters. The next three tasks involved words in Czech (the native language of participants) with the following translation to English: *lektorka* - teacher (female), *porovnat* - to compare, *nepopadnout* - to not catch. The final task involved a longer sentence : (*Tramvaj dnes uz nepojede* - The tram won't go today). Each task was recorded through a tablet as a multivariate time-serie of seven measures sampled at 150 Hz (see figure 3) : the $x$ and $y$ spatial coordinates, a time stamp, the button

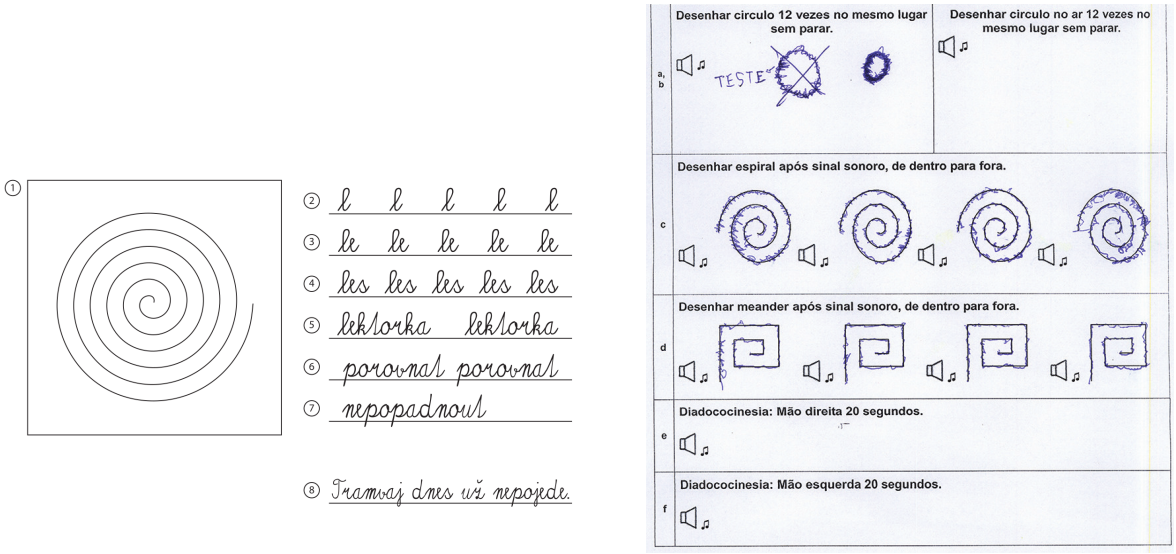


Figure 2: From Drotár et al. (2016) and Pereira et al. (2018): Examination form for the *PaHaW* (left) and *NewHandPD* (right) databases.

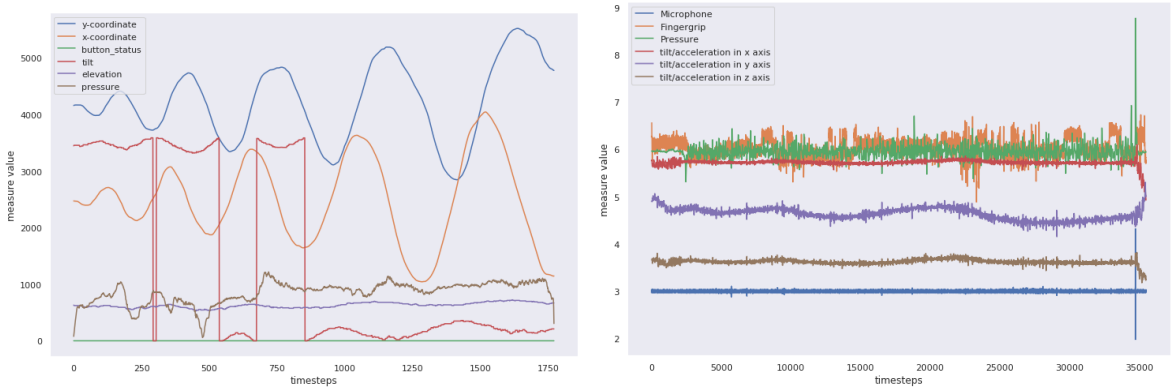


Figure 3: Spirals of the first subjects (PDs) of *PaHaW* (left) and *NewHandPD* (right). The timestamp of *PaHaW* is not plotted for a better readability. Best viewed in color.

status which indicates whether the stroke is in-air or on-paper, the pressure and, the tilt and elevation which are angles that the pen makes with $z$ and $y$ axis, respectively (Drotár et al., 2016).

The well-named *NewHandPD* database is a new version of the *HandPD* database (see appendix A.2). It comprises 66 subjects (31 PDs and 35 HCs). There are 12 tasks in the examination : 4 spirals (labeled sigSp 1, 2, 3, 4 afterwards), 4 meanders (labeled sigMea 1, 2, 3, 4 afterwards), 2 circled movements (one circle on paper, circA and another in the air, circB), and left and right-handed diadochokinesis (i.e. hand-wrist movements) : sigDiaA and sigDiaB, respectively (see Figure 2). Each task was recorded through a smart-pen as a multivariate time-serie of six measures sampled at 1000 Hz (see figure 3) : the microphone signal, the fingergrip, the pressure and the tilt/acceleration in the $x$, $y$ and $z$ axis (Pereira et al., 2018).

In the next sections we describe works using these two (and other) databases.

### 3.1.3 Hand-crafted features for machine learning

In 2016, Drotár et al. introduced novel features based on the pressure recorded by the tablet. They were inspired by established kinematic features such as *number of changes in velocity* and proposed *number of changes in pressure*. Interestingly, they obtained a better accuracy using solely these pressure features (82.5%) than using both kinematic and pressure features (81.3%) with a Support Vector Machine classifier (**SVM**, Vapnik (1998)). The same group of authors used similar features in order to assess for children's *dysgraphia* (Mekyska et al., 2016).

However, Mucha et al. (2018) outperformed Drotár et al. without using pressure features nor a **SVM** as most of the literature but the tree boosting system **XGBoost** (Chen and Guestrin, 2016) which allow them to combine different features of different tasks. They achieve 95.5% *Se*, 100% *Sp* and 97.1% accuracy which is state-of-the-art on *PaHaW*. These results are unbelievably good as their **XGBoost** system only used one feature to achieve it : the in-air horizontal velocity (median) of the sentence task. *Unbelievable* as Drotár et al. (2016) use the same feature but only attain 74.9% accuracy on the sentence task, using only the kinematic features.

In a previous work, Drotar et al. (2014) found that kinematic features from in-air movements were more discriminative than features from on-surface movements (87% vs. 78% *Se* with an **SVM** classifier). This is also confirmed by the work of Mucha et al. (see above). This is explained by Rosenblum et al. (2013b) as *"in-air time is a manifestation of* 'planning the next movement'*, as required in the sequential process of handwriting"*.

Another well-established feature used by Drotár et al. (2014); Taleb et al. (2017); Impedovo et al. (2018) is the *Shannon entropy* applied on the first few Intrinsic Mode Functions of the signal which should represent the noise of the signal (Drotár et al., 2014). Other authors use established kinematic features such as *writing velocity* and *acceleration*.

In addition to the classical PD diagnosis where they achieved 86% *Se* and 81% *Sp*, San Luciano et al. classified only early-stage PD (less than 4 years of disease duration) and obtained similar results – on their own database. This concurs with the works of Stanley et al. (2010); McLennan et al. (1972) who respectively found that spiral analysis, and, *micrographia* were early markers for PD. Furthermore, San Luciano et al. did a sensitivity analysis limited to men alone and found the *"same direction and significance of results"*.

In addition to their 2017 work, Taleb et al. (2018) achieved 94%, 92%, and 88% accuracy with an **SVM** predicting the Hoehn and Yahr stage (H&Y aka UPDRS V), UPDRS scores, and total UPDRS, respectively – on their own *PDMultiMC* database. This promises an inexpensive, non invasive and possibly remote monitoring tool for

PD progression. Although, further studies with larger databases should be used to confirm this : there are 7 stages in the modified H&Y stage (with stage 1.5 and 2.5, excluding the 0 since Taleb et al. focused on the PD only) and *PDMultiMC* contains only 5 of these : none of the PD has a 2.5 nor a 5 H&Y stage. In the same way, most of the possible UPDRS scores are not represented in the *PDMultiMC* database (as it comprises only 16 PDs).

The results of Mucha et al. on the same task are less optimistic, although they cannot be directly compared as they treat it as a regression task and evaluate their model using estimated error rate (EER) defined as :

$$\frac{1}{n \cdot r} \sum_{i=1}^{n} |y_i - \hat{y}_i| \cdot 100\% \tag{2}$$

where $y_i$ represents the true label of the $i$th observation, $\hat{y}_i$ denotes the predicted label of the $i$th observation, $n$ is the number of observations, and $r$ is the range of the values (here 4 for H&Y stage and 17 for the duration of PD). They achieve 12.51 $\pm$ 7.55 and 23.64 $\pm$ 7.55 EER for H&Y stage and PD duration tasks, respectively (average over 10-CV $\pm$ std). Notice how high is the std between the folds. They use the same **XGBoost** model described above.

All of these works are very encouraging for the design of a CDSS for PD based on handwriting examination. Moreover, early experiments suggest that the latter can be used to monitor PD progression in an inexpensive, non invasive and possibly remote way.

Although Archimedean spiral is the *most* established task for PD's assessment (see section 3.1.1), Drotár et al. (2016) found that it was the *least* discriminative (providing 10% less accuracy than *l*, *porovnat* and sentence tasks). They explain that it is because they focused on handwriting features and did not introduce spiral-specific features. This is in line with the thoughts of Phillips et al. (1991), see section 1. This motivated the use of deep learning models which are able to learn features from the raw data. We will see in the next section that the spiral task is, on the contrary, the *most* discriminative for the model of Moetesum et al. (2019).

### 3.1.4 Deep learning

Both Passos et al. (2018) and Moetesum et al. (2019) used a **CNN** pre-trained on *ImageNet*[7] (a large computer vision database) on the static images of the *HandPD* and *PaHaW* databases, respectively. Both empirically demonstrate the learning power of CNNs as they achieve 84% *Se*, 92% *Sp* and 84% *Se*, 82% *Sp* – respectively, despite the fact that the convolutional layers of the CNN they used were trained to a drastically

[7] http://www.image-net.org/

different task, namely object detection (with 1000 classes). Furthermore, both achieve competitive results with the state of the art although they only use the *static images* of the data and not the online handwriting time series (thus loosing all the *dynamic* information). Unlike Drotár et al. (2016), Moetesum et al. (2019) obtain their best results on the spiral task.

Although their work is very similar, Passos et al. and Moetesum et al. do not cite each other, one can assume that they were not aware of each other's work as their papers were published in less a year-interval.

Pereira et al. (2018) attained 97% *Se* and 94% *Sp*, using a **CNN** by transforming the *NewHandPD* online handwriting time series into a gray-scale "image" (1 row per milliseconds and 6 columns corresponding to the 6 measures, see section 3.1.2). Although their data representation creates spatial relation where there is none, they achieve incredibly good results. We believe it is partly because the *NewHandPD* database is biased because of too young HCs, see section 4.

Even though the work of Passos et al. and Moetesum et al. is impressive, by using only static images they lose all the dynamic information contained in the databases. Therefore, in section 5 we will also use a **CNN**, but, a uni-dimensional one over the online handwriting time series of *PaHaW*.

## 3.2 CDSS for PD via other examination

Many different examinations other than handwriting and symptoms have been used to build a CDSS for PD, a summary is available in table 2.

Some excellent results (above 95% *Se* and *Sp*) have been achieved using vocal analysis by extracting hand-crafted features (e.g. fundamental frequency), see Chen et al. (2016); Tsanas et al. (2012); Zuo et al. (2013). Most works cited in table 2 use hand-crafted features coupled with an **SVM**. Although Oh et al. (2018) used an uni-dimensional CNN over EEG data which inspired us, see section 5.1.

Moreover, Raza et al. (2017) performed a pilot study where their CDSS was actually used on new patients. They studied *tremor* via an examination where subjects had to stretch their arms (data was collected through an accelerometer and a gyroscope). Their task was a sort of differential diagnosis between PD and other movement disorders – namely Vascular pseudo-parkinsonism, Alzheimer's disease, Dystonia and Benign Essential Tremor – merged as one class. Their CDSS achieved 90% *Se* and 60% *Sp* on their original test set (using Google's Prediction API, a Machine learning black-box) and 80% *Se* and 75% *Sp* on new patients. Although this study has been conducted on very small data sets, these results are encouraging for the clinical use of CDSS.

Table 2: Summary of the different examinations (except handwriting) or studied symptom used in CDSS for PD.

| **Examination or studied symptom** | **Authors** |
|---|---|
| Vocal | Chen et al. (2016); Tsanas et al. (2012); Zuo et al. (2013) |
| Exhaled Breath | Tisch et al. (2013) |
| Electromyography (EMG) | Loconsole et al. (2019); Arvind et al. (2010) |
| Electroencephalography (EEG) | Oh et al. (2018); Ruffini et al. (2016) |
| Magnetic Resonance Imaging (MRI) | Rana et al. (2015) |
| Gait | Zhao et al. (2018); Abdulhay et al. (2018) |
| Tremor | Raza et al. (2017) |
| Bradykinesia | Eskofier et al. (2016) |

# 4 Choice of database

In this section we motivate our choice to work on the *PaHaW* over the *NewHandPD* database. We first introduce some background materials about PD and its relation with aging, then, we will analyze the two databases before concluding.

## 4.1 PD and aging

> *"Handwriting deterioration in Parkinson patients correlates both with their age and the severity of their disease."*
>
> — Walton (1997)

Aging is the biggest risk factor for developing PD (Reeve et al., 2014). Moreover, several symptoms of PD are shared with normal aging, such as :

- Loss of dopaminergic neurons (Rodriguez et al., 2015). It even seems to correlate with motor performance as much in both HCs and PDs (Pujol et al., 1992).
- *Bradykinesia* (Mortimer, 1988). Rosenblum et al. (2013a) found that age accounted for 32% of the average stroke in-air time.
- *Tremor* (Walton, 1997).

However, both *NewHandPD* and *PaHaW* have older PDs than HCs : 14 and 7 years in average, respectively. We will study the significance of these gaps in the next section. We will start by analyzing the discriminative power of the duration of the task, then we will interpret it as a consequence of the subjects' age.

## 4.2 Data analysis and classification

### 4.2.1 Task duration

As we mentioned in the previous section, PDs suffer from *bradykinesia*, which causes them to complete a graphomotor task in more time than usually required. However, we can observe a great difference between the *NewHandPD* and the *PaHaW* databases, see figure 4 :

- on *NewHandPD*, we can clearly see a threshold on the duration axis which separates most of PDs and HCs.
- on *PaHaW*, however, the durations are more similar between HCs and PDs.

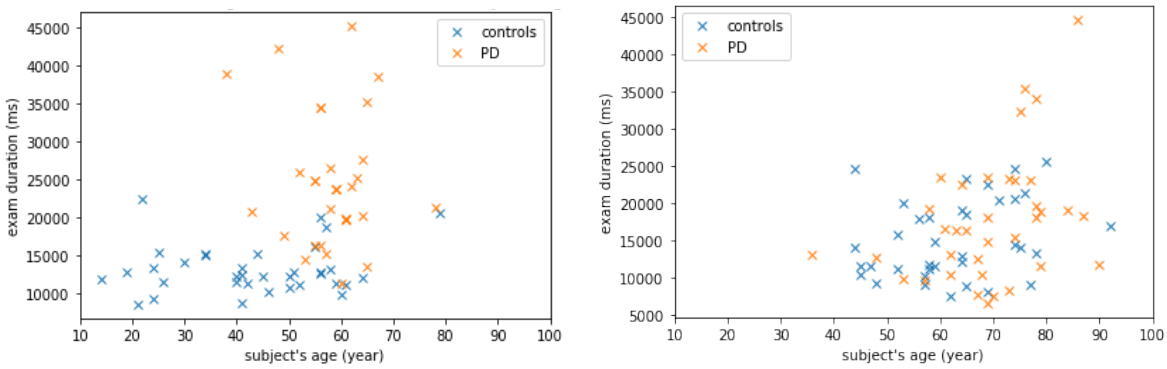


Figure 4: Average duration of tasks vs. subject's age. Left : *NewHandPD*, right : *PaHaW*.

This phenomenon is observable in every single tasks, we applied Student's t-test to study it (see section 2.2). Although sigDiaA task was excluded because it failed Bartlett's test and SigMea 1, 2, 4 and SigSp 2, 3 and 4 were excluded because either the HC or the PD group failed the Shapiro-Wilk's test ; all the other tasks of *NewHandPD* have a strongly significant difference between HCs and PDs' task duration ($p < 2^{-6}$).

On *PaHaW* however, although the *l* task was excluded because it failed the Bartlett's test, none of the task have a significant difference between HCs and PDs' task duration ($p > 0.1$).

To further confirm these findings, we used the scikit-learn[8] implementation of LDA (see section 2.3). The model is fed the duration of the task as sole feature. See results in table 3. In order to have comparable results with Afonso et al. (2017), random 50-50 train-test split with 15 runs was used to evaluate the model (see section 2.3.4). The evaluation method of Pereira et al. (2018) is unknown.

The duration of the four meanders and the four spirals is averaged to provide the results of the lines *meanders* and *spirals*, respectively. In the same way, in *all tasks* the duration of all tasks are averaged. Majority voting provided similar performance. We

[8]https://scikit-learn.org/

Table 3: Accuracy (%) depending on the task (average over the 15 runs ± std). The best results are print in bold.

| **task** | **LDA (our model)** | **Afonso et al.** | **Pereira et al.** |
|---|---|---|---|
| circA | **80.61** ± 4.40 | 76.17 ± 6.92 | 68.04 ± 2.96 |
| circB | **83.43** ± 4.68 | 76.69 ± 5.38 | 73.41 ± 3.66 |
| sigDiaA | **75.96** ± 5.90 | 68.69 ± 7.26 | 73.59 ± 3.57 |
| sigDiaB | **77.17** ± 5.52 | 66.30 ± 7.38 | 76.32 ± 5.18 |
| meanders | **81.82** ± 3.13 | 81.07 ± 2.60 | 80.75 ± 2.08 |
| spirals | **82.02** ± 3.02 | 81.03 ± 2.40 | 78.26 ± 1.97 |
| all tasks | 83.84 ± 3.26 | - | **95.74** ± 1.60 |

can see that it does not improve performance, this is because the same subjects are misclassified over every tasks since we use only one feature which is very dependent on the subject.

For comparison purposes, we display in table 3 the results of Afonso et al. (2017) when using first OPF then SVM because it is their bests. In the same way the results of Pereira et al. (2018) are where they used the *"CNN-ImageNet"* model on *"images"* of size ($128 \times 128$) because it is their bests. Cf. to their respective papers for further details. Notice how we outperform both of them on every single task assessment. However, Pereira et al. achieved better results after combining their predictions over all tasks using majority voting.

The same study was performed on *PaHaW* where Drotár et al. (2016) outperform our model by a large margin on all tasks (between 11% and 25% accuracy) as it barely gets above chance level on some tasks and falls behind on others.

This study motivates the use of *PaHaW* over *NewHandPD*, we will aim at explaining these findings in the next section.

### 4.2.2 Influence of subjects' age

As mentioned in section 4.1, not only PD is correlated with age but several symptoms are shared among PD and normal aging, including *bradykinesia* which we have observed in the previous section thanks to the task duration.

This is very clear when looking at figure 4, there is an obvious correlation between the subjects' ages and their average task duration in both databases (confirmed by Pearson's correlation coefficient). Moreover, in *NewHandPD*, most HCs and PDs are grouped into two clusters, unlike in *PaHaW*. This is because the latter is more balanced in regards to its subjects' age.

However, we can still observe a tendency of younger HCs in *PaHaW*, thus when applying a t-test (as in the previous section), there is a significant difference between the PDs and the HCs' age ($p = 1.3 \times 10^{-2}$). In *NewHandPD* the difference is statistically stronger ($p = 8 \times 10^{-6}$).

To further confirm these findings, as in the previous section we classified PDs and HCs using LDA, this time using subjects' age as sole feature. The same evaluation method was applied (i.e. random 50-50 train-test split with 15 runs) and a classification accuracy of $75.76 \pm 3.67$ and $63.70 \pm 7.49$ was obtained on *NewHandPD* and *PaHaW*, respectively.

These results concur with the statistical analysis and suggests that the *NewHandPD* is strongly biased while the *PaHaW* database is slightly biased.

## 4.3 Conclusion

Both of the *NewHandPD* and *PaHaW* databases have younger HCs than PDs, even though PD share symptoms with normal aging (see section 4.1). This allows to classify subjects based on their age with an accuracy significantly higher than chance level on both databases. However, our model reaches 76% accuracy on *NewHandPD* and only 64% accuracy on *PaHaW* using this feature.

Moreover, as PD and elderly people suffer from *bradykinesia*, we are able to outperform state of the art on *NewHandPD*, using the task duration as sole feature. However, this is not true on *PaHaW* where our model barely reaches chance level using the same feature.

For these reasons we decided to work on the *PaHaW* database in the next section.

# 5 Machine learning

This section describes the work related to machine learning done during my internship. We start by presenting the proposed model architectures and report the training process, then, we analyze the performance of our models and compare it to the related works. We then go over the challenges and difficulties encountered during this work.

## 5.1 Proposed architectures

For time series data as *PaHaW*, Recurrent Neural Networks (**RNNs**) such as **Long Short Term Memory** are the usual go-to tool (see, e.g. Salehinejad et al. (2017)), and, even though they are not state-of-the-art (Keysers et al., 2016), they were successfully applied to several online handwriting tasks (see, e.g. Graves et al. (2008) who worked on online handwriting recognition).

For these reasons, we first focused on **RNNs** at the beginning of my internship. However, we quickly faced one of the biggest challenge of our work : the dimensionality of the *PaHaW* dataset. Indeed, the *PaHaW* data (see section 3.1.2) is sampled at 150 Hz, thus, handwriting tasks consists of several *thousands* of timesteps. It is very difficult for RNNs to capture such long-term dependencies (Bai et al., 2018). In addition, the *PaHaW* database comprises only *72* subjects whereas the successfully applied RNNs required gigantic amount of data (Salehinejad et al., 2017).

Bai et al. empirically compare CNNs to RNNs on several sequence modeling tasks and found that CNNs exhibited a longer memory than RNNs. To allow the model for a large scope the authors use dilated convolutions. Dilated convolutions (aka *à trous* algorithm, Holschneider et al.) add space between the kernel neurons. Moreover, Oh et al. (2018) (see section 3.2) who also work on a very small PD dataset consisting of (very long) EEG signals successfully use CNNs by subsampling the EEG signals into windows of 256 timesteps. This technique augments the dataset and a similar approach was used in section 5.1.3.

The next sections present the different proposed models – namely **Baseline** and **StrokeCNN** – that are compared with state of the art in section 5.2.2. All the models share the same training process which is described in section 5.1.1 and the basic CNN architecture (see section 2.3.2). The models were implemented using the PyTorch framework[9].

In order to find the best hyperparameters for the different models, a random search was performed in the following sets :

- $num\_layers : \{1, 2, 3, 4\}$. Number of *Conv-Pool* blocks (see table 5).

[9] https://pytorch.org/

- *learning_rate* : $\{0.001, 0.0001\}$. Used to optimize the weights of the model during SGD.
- *dropout* : $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Applied after every pooling layers to regularize the model.
- *hidden_size* : $\{4, 8, 16, 32\}$. Number of kernels.
- *dilation* : $\{1, 2, 4, 8, 16\}$. Dilation of the convolution.
- *conv_kernel* : $\{1, 2, 4, 8, 16\}$. Size of the convolutional kernel.
- *pool_kernel* : $\{8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096\}$. Size of the pooling kernel.

The models were evaluated using 10-fold cross-validation (see section 2.3.4) in order to have comparable results with Drotár et al.; Mucha et al.; Moetesum et al..

Next section describes the training process which is common to every model.

### 5.1.1 Training process

In order to facilitate the training (LeCun et al., 2012), each data sample was standardized independently so that each measure-vector $X$ (e.g. $x$-coordinate) has a mean of 0 and a std of 1. This is done by subtracting the average $\bar{X}$ and dividing the std $\sigma$ of each measure-vector $X$ as this :

$$X \leftarrow \frac{X - \bar{X}}{\sigma} \tag{3}$$

The feature map extracted by the convolutional layers should have the same dimension for all samples so that the FC layer would be able to learn the optimal weight of each feature. In order to achieve that, we can apply two non-exclusive methods :

1. trim and/or pad all the samples to a given length (done in **StrokeCNN**).
2. set the pooling size to the maximum length of the samples so that each kernel only outputs one feature (done in **Baseline**).

Every convolutional layer has a Rectified Linear Unit (ReLU) activation which is defined as :

$$\text{ReLU}(x) = max(0, x) \tag{4}$$

The model is regularized by applying dropout after the pooling layers. Dropout helps to prevent neural networks from overfitting the training set by randomly dropping neurons during the training (Srivastava et al., 2014).

The feature map extracted by the convolutional layers is then flattened and fed to a FC layer with a single neuron which is used for binary classification after applying a Sigmoid activation. The Sigmoid function is is defined as :

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

The predictions of the model are saved after each epoch in order to allow for majority voting as if *eight* models were trained on all *eight* tasks simultaneously (see section 2.3.3).

We train the model using Adam optimizer (Kingma and Ba, 2014) which is based on SGD. The optimizer minimizes the binary cross entropy loss function which is defined as :

$$L(y, \hat{y}) = y \log(f(\hat{y})) - (1 - y) \log(1 - f(\hat{y})) \tag{6}$$

Where $y$ is the label (aka target) of a data sample, $\hat{y}$ is the prediction of the model and $f$ is the, above defined, Sigmoid function.

In order to speed up the training process, weight normalization (Salimans and Kingma, 2016) is applied to the weights $\mathbf{w}$ of the convolutional layers. Those weights are reparameterized by replacing $\mathbf{w}$ by $g\frac{\mathbf{v}}{\|\mathbf{v}\|}$ before each forward pass, thus giving $\|\mathbf{w}\| = g$. The parameters $g$ and $\mathbf{v}$ are then optimized using gradient descent.

The next sections present the different proposed models – namely **Baseline** and **StrokeCNN** – that are compared with state of the art in section 5.2.2.

### 5.1.2 Baseline

The **Baseline** model's architecture is described in table 4. A dropout of 0.2 is applied after the *Pool* layer.

This model is trained – with a learning rate of $10^{-3}$ – on the whole task, unlike **StrokeCNN**. The random search provided some better architectures that we will not present as they fell behind **StrokeCNN**, and, **Baseline** is interesting as it does not take the sequential aspect of the data into account (because its kernels are of size 1 and the feature map is pooled to a length of 1).

Notice how small is this model : it only consists of *451* parameters. Larger models

Table 4: **Baseline** architecture. *For the spiral task. Length of the input (16071 here) is dependent on the task, cf. appendix A.3.

| **Name** | **Input shape*** | **Kernel size** | **Number of kernels** | **Activation** | **Output shape** |
|---|---|---|---|---|---|
| *Conv* | $(16071 \times 7)$ | 1 | 50 | ReLU | $(16071 \times 50)$ |
| *Pool* | $(16071 \times 50)$ | 16071 | 50 | - | $(1 \times 50)$ |
| *FC* | (50) | - | - | Sigmoid | (1) |

Table 5: **StrokeCNN** architecture. *For the $l$ task. Length of the input (752 here) is dependent on the task, cf. appendix A.3.

| Name | Input shape* | Kernel size | Dilation | Number of kernels | Activation | Output shape |
|---|---|---|---|---|---|---|
| $Conv_1$ | $(752 \times 7)$ | 16 | 16 | 128 | ReLU | $(512 \times 128)$ |
| $Pool_1$ | $(512 \times 128)$ | 8 | - | 128 | - | $(64 \times 128)$ |
| $Conv_2$ | $(64 \times 128)$ | 4 | 16 | 32 | ReLU | $(16 \times 32)$ |
| $Pool_2$ | $(16 \times 32)$ | 16 | - | 32 | - | $(1 \times 32)$ |
| $Conv_3$ | $(1 \times 32)$ | 1 | - | 128 | ReLU | $(1 \times 128)$ |
| $Pool_3$ | $(1 \times 128)$ | 1 | - | 128 | - | $(1 \times 128)$ |
| $FC$ | (128) | - | - | - | Sigmoid | (1) |

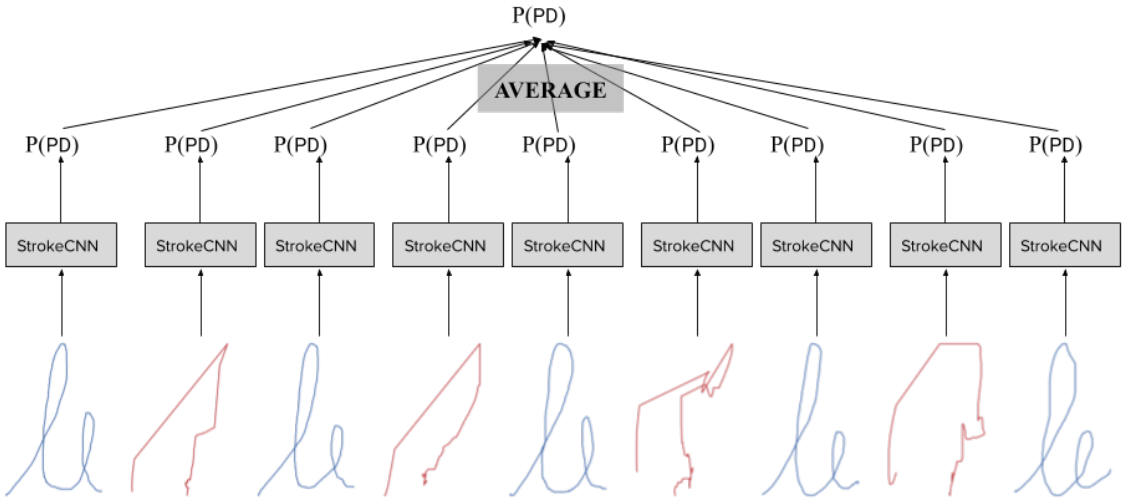


Figure 5: **StrokeCNN** stroke fusion scheme. Note that the data has been standardized and that each stroke is displayed at its own scale.

consistently overfitted the data. The technique presented in the next section allowed for larger models.

### 5.1.3 StrokeCNN

The architecture of **StrokeCNN** is described in table 5. A dropout of 0.3 is applied after all pooling layers. Note how strange is the last convolutional layer as it transforms 32 feature maps of size 1 into 128 feature maps of size 1. Removing it provided similar performance.

Unlike **Baseline** which is trained on the whole task, **StrokeCNN** is trained – with a learning rate of $10^{-4}$ – on all strokes of a given task independently (see figure 5). The final prediction for a given task (i.e. subject) is the average of its strokes' predictions. This was done in order to augment the dataset and we will see in section 5.2.2 that it significantly improves the model's performance. This technique allowed for larger networks than **Baseline** : **StrokeCNN** has *78 times* more parameters.

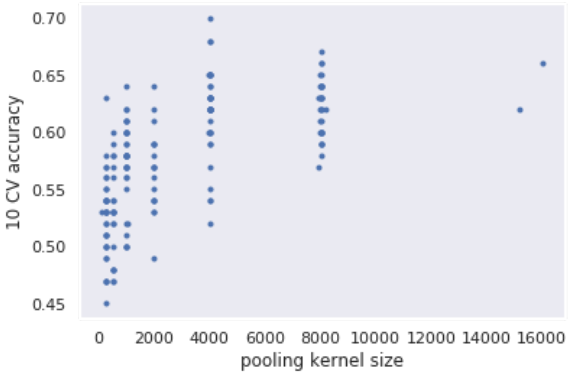


Figure 6: Accuracy vs. *pool_kernel* over 170 experiments on the spiral task.

Table 6: Accuracy (average over the 10 folds in % ± std) compared to Drotár et al., Moetesum et al. and Mucha et al.. Best results are print in bold.

| **task / model** | **Baseline** | **StrokeCNN** | **Drotár et al.** | **Moetesum et al.** | **Mucha et al.** |
|---|---|---|---|---|---|
| spiral | 67 ± 15 | 56 ± 16 | 63 | **76** ± 8.0 | - |
| *l* | 60 ± 12 | **80** ± 12 | 72 | 62 ± 8.0 | - |
| *le* | 62 ± 17 | 70 ± 17 | **71** | 57 ± 9.0 | - |
| *les* | 60 ± 12 | **70** ± 11 | 66 | 60 ± 8.0 | - |
| *lektorka* | 60 ± 20 | 59 ± 14 | **65** | 60 ± 7.0 | - |
| *porovnat* | 62 ± 15 | 66 ± 16 | **73** | 51 ± 9.0 | - |
| *nepopadnout* | 58 ± 16 | **71** ± 16 | 68 | 68 ± 0.0 | - |
| sentence | 63 ± 09 | 65 ± 21 | **76** | 51 ± 8.0 | - |
| all tasks | 60 ± 13 | 78 ± 13 | 81 | 83 ± 9.0 | **97** ± 5.7 |

## 5.2 Analysis of results

### 5.2.1 CNN hyperparameters

Early experiments suggested that a learning rate of $10^{-3}$ or $10^{-4}$ was suited for most models, this was confirmed by all the random search experiments. This is in line with the findings of Greff et al. (2016) on **Long Short Term Memory** which observed that the learning rate is dependent on the dataset but that, independently of the dataset, there is a large basin (up to two orders of magnitude) of good learning rates inside of which the performance does not vary much.

A random search over the spiral task revealed that, when fixing *num_layers* at 1, the *pool_kernel* should be quite large, see figure 6.

Making similar observations for the other hyperparameters is difficult as there is a lot of interaction between those (e.g. between *conv_kernel* and *dilation*).

### 5.2.2 Empirical comparison

The results are summarized in table 6.

It is interesting that we achieve decent results with **Baseline** despite the fact that it does not take the sequential aspect of the data into account (see section 5.1.2). We will see in the next section that the model extracts basic statistical functionals such as *minimum* and *maximum*. However **StrokeCNN** outperforms **Baseline** on most tasks and by a large margin after majority voting. **StrokeCNN** falls below **Baseline** on the spiral and the *lektorka* task which strokes are highly variable in length. The model performance was found to be negatively correlated with the std of the strokes length (p-value $< 0.01$). Unlike **Baseline**, **StrokeCNN** has a very large scope : the first convolutional layer has a kernel of 16 with a dilation of 16 meaning it has a scope of $16 \times (16 - 1) + 1 = 241$ timesteps.

We achieve competitive results with Drotár et al. and Moetesum et al. using **Stroke-CNN**. These results are very encouraging for end-to-end learning on a small dataset like ours. However, we fall far below Mucha et al. who use the tree boosting system **XGBoost** which allows them to combine different features of different tasks (see section 3.1).

### 5.2.3 Explainability of the models

Unfortunately, since our models are uni-dimensional and were trained on time series and not images, it is quite difficult to interpret the *weights* of the convolutional layers by visualizing them as we are used to see in computer vision papers. However, we're able to interpret the *features* extracted by the model. As **Baseline** is simpler then **StrokeCNN** it is easier to interpret : as one could expect, since **Baseline** has a kernel size of 1, it mostly monitors peaks in the measures (i.e. a sudden drop or rise in the measures' value). Some features also seem to be related to several measures, see figure 7.

**StrokeCNN** is less straightforward as the first convolutional layer has a kernel of

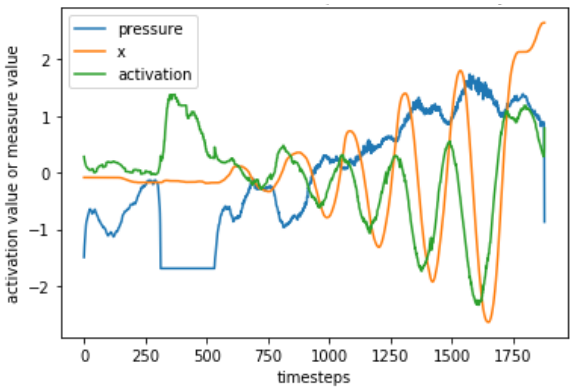


Figure 7: Feature #4 extracted by **Baseline** (before *ReLU*) plotted along subject #40 (HC)'s spiral's *pressure* and *x-coordinate*. The feature first activates when there's a drop in the pressure but then seems to follow the x coordinate. Best viewed in color.

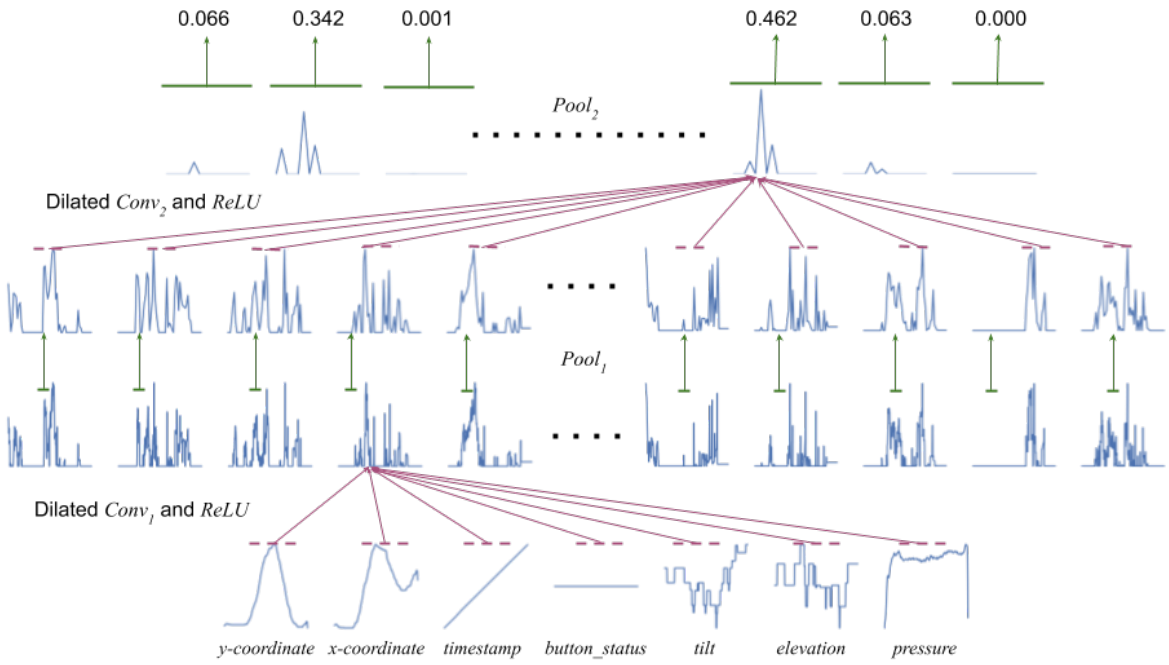


Figure 8: First $l$ stroke of subject #8 (PD) as seen by **StrokeCNN** after 40 epochs of training.

16 with a dilation of 16 giving it a scope larger than the average stroke length, and, the second convolutional layer provides abstract features, see figure 8.

The lack of explainability of the models is one of the biggest limitations of our work. On the contrary, the model of Mucha et al. provides interesting insights for practitioners as it focus on a very few, hand-crafted (thus understandable) features.

## 5.3 Challenges and difficulties

As mentioned in section 5.1, one of the biggest challenges of our work was the dimensionality of the dataset. Indeed, the *PaHaW* data (see section 3.1.2) is sampled at 150 Hz, thus, handwriting tasks consists of several *thousands* of timesteps. In addition, the *PaHaW* database comprises only *72* subjects whereas most deep learning models require gigantic amount of data (LeCun et al., 2015). This goes without saying as it needs to learn features from the raw data, thus, a large amount of it is necessary to be able to generalize.

This inhibited the use of **RNNs** which are the usual go-to models for time series (see section 5.1). Moreover, it hindered the generalization capacity of the model : even when using a very small network compared to deep learning standards we still heavily overfit the training set.

Furthermore, the smallness of the dataset inhibited a very common practice in deep learning : early stopping. Early stopping consists in having a validation set in addition to the training and test set. The validation set is then used to evaluate the model and

stop the training on a given criteria to prevent overfitting. However as the *PaHaW* dataset is very small, the test set is smaller, hence, the performance of the model greatly varies between folds (this is testified by the high std in table 6).

In order to solve this challenge, we split the tasks into strokes in order to augment the dataset (see section 5.1.3), this significantly improved the performance of the model. Other data augmentation techniques (translating, rotating, speeding up, slowing down, clipping, window warping, flipping and re-scaling) have not been successful. This is because a key concept of data augmentation is that the deformations applied to the labeled data do not change the semantic meaning of the labels (Salamon and Bello, 2017). However this key concept is not as straightforward in our case as in, e.g. automatic speech recognition or computer vision : a rapidly drawn spiral is still a spiral but is it still a spiral drawn by a PD ?

Another – more drastic – solution would have been to work on another, larger, dataset or use transfer learning. However, the only large PD handwriting database to our knowledge is that of San Luciano et al.. Unfortunately, this database is not available publicly and the authors did not reply to our solicitation. We further discuss transfer learning in the next section.

# 6 Conclusion

## 6.1 Contributions

To the best of our knowledge, this work is the first to learn PD-discriminative features in an end-to-end manner from the – almost standard – *PaHaW* database. If several of our experiments have failed (see section 5.3), it is still interesting to know they have been carried out. Moreover, splitting the handwriting into strokes has allowed for competitive results with the literature and confirms that this method is a valid data augmentation technique (see section 5.1.3).

Every piece of code implemented during this internship has been made available on GitHub[10] under the – very permissive – MIT License[11]. The code is fully-documented and allows for the reproducibility of our results.

A short paper was submitted to the fourth edition of the Paris-Saclay Junior Conference on Data Science and Engineering (JDSE)[12]. The paper was accepted and we will present it through a poster (Lerner and Likforman-Sulem, 2019). JDSE is addressed to junior scientists such as first year PhDs or Master students of *Université Paris-Saclay* or *Institut Polytechnique de Paris*. The conference will be held on September 12th and 13th at CentraleSupélec, Paris-Saclay campus.

As my defense at *Université Paris Descartes* and JDSE should provide interesting feedbacks, we plan to submit a full-length paper for the ninth International Conference on Pattern Recognition Applications and Methods (ICPRAM)[13]. The conference will be held on February 22-24th in Malta, Italy.

## 6.2 Limitations

One of the biggest limitation of our work is the lack of explainability of our models (see section 5.2.3). On the contrary, the model of Mucha et al. provides interesting insights for practitioners as it focus on a very few, hand-crafted (thus understandable) features. Explainability of deep learning models is an active research field and several journals call for papers related to it[14,15].

Moreover, we share numerous limitations with the rest of the literature, including several that are not mentioned in it.

PDs of the *PaHaW* database are under medication (L-dopa COMT inhibitor and/or

[10] https://github.com/PaulLerner/deep_parkinson_handwriting
[11] https://opensource.org/licenses/MIT
[12] https://jdse-paris.github.io/jDSE2019/
[13] http://www.icpram.org/
[14] https://www.journals.elsevier.com/pattern-recognition/call-for-papers/call-for-paper-on-special-issue-on-explainable-deep-learning
[15] https://www.journals.elsevier.com/artificial-intelligence/call-for-papers/special-issue-on-explainable-artificial-intelligence

a dopamine agonist (Drotár et al.)). Even though this can make the classification task harder as *"treated PD subjects may closely resemble controls"* (San Luciano et al.) it can also introduce bias. Indeed, excess of L-dopa may cause *dyskinesia* and Zham et al. have excluded several of their subjects because of this. Drotár et al. do not go over in detail about this. Moreover, training a model over medicated PDs raises another question : will the model be able to diagnose new, drug-free subjects ? These questions could be answered using the database of San Luciano et al. which comprises 138 PDs among which 46% are medicated. Unfortunately, this database is not available publicly and the authors did not reply to our solicitation.

Another limitation is the universality of the examination, indeed, Mucha et al.; Drotár et al. and ourselves' best results are based on horizontal writings. However, symptoms such as *micrographia* may be related to the writing direction (Thomas et al., 2017) and the findings of Mucha et al. support that it also affects the kinematics of the handwriting.

Finally, the works described in this report are focused on the classification of PDs and HCs. Drotár et al. (2016) state that the probability of an inaccurate diagnosis using classical clinical methods is approximately 25%, referring to Hughes et al. (2002). However this number refers to a differential diagnosis between PD and other parkinsonian syndromes, namely multiple system atrophy and progressive supranuclear palsy. By the way, Hughes et al. report a $Se$ of 91% and a $Sp$ of 98% for PD. I was not able to find any classical clinical study that report diagnosis accuracy between PDs and HCs. Knowing whether handwriting examination is able to discriminate PDs and HCs, and, PDs and other neurodegenerative diseases are two different questions and the latter is up to debate : Yu et al. (2017) reported significant differences in handwriting kinematics between patients with PD and essential tremor, however, Ling et al. (2012) show that *micrographia* is more frequent in progressive supranuclear palsy than in PD and that finger tap analysis is a better tool to distinguish the two diseases.

Therefore, I believe that the works presented until now can take three different directions :

1. differential CDSS between PD and other neurodegenerative diseases.
2. monitoring of PD's progression (works of Taleb et al.; Mucha et al. are promising).
3. early detection of PD (San Luciano et al. reported very encouraging results).

## 6.3 Future Works

Several perspectives to address the limitations of our work are described in the previous section. Here we focus on methods which would improve the performance of our model

(described in section 5.1).

Section 5.3 goes over the difficulties encountered during this work, in a nutshell : deep learning models require large amount of data. In the absence of larger PD handwriting database, one could apply transfer learning. Transfer learning was succesfully applied by Moetesum et al.; Passos et al., see section 3.1.4. Questions are then : what the model should be trained to first, and, from which data ? The latter sounds like common sense but answers may be multiple. We believe, however, that the model should be pre-trained on online handwriting data. The former question is more tricky, we propose two possible answers : auto-encoding and mode detection.

Auto-encoding is an unsupervised learning task in which the model needs to reconstruct its input. The architecture for convolutional auto encoders is then similar in that depicted in section 2.3.2, only instead of the FC layers are deconvolutional layers which reconstruct the input from the feature maps. The idea is that the model learns relevant features as they allow for this reconstruction. An advantage of this task is that it is unsupervised, thus, the model can be trained on large, unlabelled datasets. However, in low-dimensional data such as online handwriting, the model might learn the identity function without learning relevant features specific to the data. Nonetheless, auto-encoding has been succesfully applied to online handwriting tasks, see Fayyaz et al. who work on signature verification.

Mode detection consists in detecting if the handwriting is a drawing or a text. The IAMonDo-database collected by Indermühle et al. allows for it as it contains tens of thousands of words and thousands of drawings produced by 200 writers. We hypothesize that the model would learn interesting features from this task, however, this is *terra incognita*.

Early experiments on both tasks provided encouraging results.

# References

Abdulhay, E., Arunkumar, N., Narasimhan, K., Vellaiappan, E., Venkatraman, V., 2018. Gait and tremor investigation using machine learning techniques for the diagnosis of parkinson disease. Future Generation Computer Systems 83, 366–373.

Afonso, L.C.S., Pereira, C.R., Weber, S.A.T., Hook, C., Papa, J.P., 2017. Parkinson's disease identification through deep optimum-path forest clustering, in: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE. pp. 163–169.

Altman, D.G., Bland, J.M., 1994. Diagnostic tests. 1: Sensitivity and specificity. BMJ: British Medical Journal 308, 1552.

Arvind, R., Karthik, B., Sriraam, N., Kannan, J.K., 2010. Automated detection of pd resting tremor using psd with recurrent neural network classifier, in: 2010 International Conference on Advances in Recent Technologies in Communication and Computing, IEEE. pp. 414–417.

Bai, S., Kolter, J.Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. ArXiv abs/1803.01271, 1.

Berardelli, A., Rothwell, J., Thompson, P., Hallett, M., 2001. Pathophysiology of bradykinesia in parkinson's disease. Brain 124, 2131–2146.

Bidet-Ildei, C., Pollak, P., Kandel, S., Fraix, V., Orliaguet, J.P., 2011. Handwriting in patients with parkinson disease: Effect of l-dopa and stimulation of the sub-thalamic nucleus on motor anticipation. Human movement science 30, 783–791.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L., Lijmer, J.G., Moher, D., Rennie, D., De Vet, H.C., et al., 2015. Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies. Radiology 277, 826–832.

Chen, H.L., Wang, G., Ma, C., Cai, Z.N., Liu, W.B., Wang, S.J., 2016. An efficient hybrid kernel extreme learning machine approach for early diagnosis of parkinson's disease. Neurocomputing 184, 131–144.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM. pp. 785–794.

Drotar, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., Faundez-Zanuy, M., 2014. Analysis of in-air movement in handwriting: A novel marker for parkinson's disease. Computer methods and programs in biomedicine 117, 405–411.

Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., Faundez-Zanuy, M., 2014. Decision support framework for parkinson's disease based on novel handwriting markers. IEEE Transactions on Neural Systems and Rehabilitation Engineering 23, 508–516.

Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., Faundez-Zanuy, M., 2016. Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease. Artificial intelligence in Medicine 67, 39–46.

Eskofier, B.M., Lee, S.I., Daneault, J.F., Golabchi, F.N., Ferreira-Carvalho, G., Vergara-Diaz, G., Sapienza, S., Costante, G., Klucken, J., Kautz, T., et al., 2016. Recent machine learning advancements in sensor-based mobility analysis: Deep learning for parkinson's disease assessment, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 655–658.

Fayyaz, M., Hajizadeh_Saffar, M., Sabokrou, M., Fathy, M., 2015. Feature representation for online signature verification. arXiv preprint arXiv:1505.08153 .

Goetz, C.G., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stebbins, G.T., Stern, M.B., Tilley, B.C., Dodel, R., Dubois, B., et al., 2007. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): process, format, and clinimetric testing plan. Movement Disorders 22, 41–47.

Graça, R., e Castro, R.S., Cevada, J., 2014. Parkdetect: Early diagnosing parkinson's disease, in: 2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA), IEEE. pp. 1–6.

Graves, A., Liwicki, M., Bunke, H., Schmidhuber, J., Fernández, S., 2008. Unconstrained on-line handwriting recognition with recurrent neural networks, in: Advances in neural information processing systems, pp. 577–584.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J., 2016. Lstm: A search space odyssey. IEEE transactions on neural networks and learning systems 28, 2222–2232.

Holschneider, M., Kronland-Martinet, R., Morlet, J., Tchamitchian, P., 1990. A real-time algorithm for signal analysis with the help of the wavelet transform, in: Wavelets. Springer, pp. 286–297.

Hubel, D.H., Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology 160, 106–154.

Hughes, A.J., Daniel, S.E., Ben-Shlomo, Y., Lees, A.J., 2002. The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. Brain 125, 861–870.

Impedovo, D., Pirlo, G., Vessio, G., 2018. Dynamic handwriting analysis for supporting earlier parkinson's disease diagnosis. Information 9, 247.

Indermühle, E., Frinken, V., Bunke, H., 2012. Mode detection in online handwritten documents using blstm neural networks, in: 2012 International Conference on Frontiers in Handwriting Recognition, IEEE. pp. 302–307.

Isenkul, M., Sakar, B., Kursun, O., 2014. Improved spiral test using digitized graphics tablet for monitoring parkinson's disease, in: Proc. of the Int'l Conf. on e-Health and Telemedicine, pp. 171–5.

Keysers, D., Deselaers, T., Rowley, H.A., Wang, L.L., Carbune, V., 2016. Multi-language online handwriting recognition. IEEE transactions on pattern analysis and machine intelligence 39, 1180–1194.

Kim, Y., 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 abs/1408.5882, 1.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.R., 2012. Efficient backprop, in: Neural networks: Tricks of the trade. Springer, pp. 9–48.

Lerner, P., Likforman-Sulem, L., 2019. Classification of online handwriting time series for parkinson's disease diagnosis using deep learning, in: 2019 Junior Conference on Data Science and Engineering (JDSE).

Letanneux, A., Danna, J., Velay, J.L., Viallet, F., Pinto, S., 2014. From micrographia to parkinson's disease dysgraphia. Movement Disorders 29, 1467–1475.

Likforman-Sulem, L., Esposito, A., Faundez-Zanuy, M., Clémençon, S., Cordasco, G., 2017. Emothaw: A novel database for emotional state recognition from handwriting and drawing. IEEE Transactions on Human-Machine Systems 47, 273–284.

Ling, H., Massey, L.A., Lees, A.J., Brown, P., Day, B.L., 2012. Hypokinesia without decrement distinguishes progressive supranuclear palsy from parkinson's disease. Brain 135, 1141–1153.

Loconsole, C., Cascarano, G.D., Brunetti, A., Trotta, G.F., Losavio, G., Bevilacqua, V., Di Sciascio, E., 2019. A model-free technique based on computer vision and semg for classification in parkinson's disease by using computer-assisted handwriting analysis. Pattern Recognition Letters 121, 28–36.

McLennan, J., Nakano, K., Tyler, H., Schwab, R., 1972. Micrographia in parkinson's disease. Journal of the neurological sciences 15, 141–152.

Mekyska, J., Faundez-Zanuy, M., Mzourek, Z., Galaz, Z., Smekal, Z., Rosenblum, S., 2016. Identification and rating of developmental dysgraphia by handwriting analysis. IEEE Transactions on Human-Machine Systems 47, 235–248.

Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T., 2017. Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics 19, 1236–1246.

Moetesum, M., Siddiqi, I., Vincent, N., Cloppet, F., 2019. Assessing visual attributes of handwriting for prediction of neurological disorders—a case study on parkinson's disease. Pattern Recognition Letters 121, 19–27.

Mortimer, J.A., 1988. Human motor behavior and aging. Annals of the New York Academy of Sciences 515, 54–66.

Mucha, J., Mekyska, J., Galaz, Z., Faundez-Zanuy, M., Lopez-de Ipina, K., Zvoncak, V., Kiska, T., Smekal, Z., Brabenec, L., Rektorova, I., 2018. Identification and monitoring of parkinson's disease dysgraphia based on fractional-order derivatives of online handwriting. Applied Sciences 8, 2566.

Oh, S.L., Hagiwara, Y., U, R., RAJAMANICKAM, Y., , A., Acharya, U.R., 2018. A deep learning approach for parkinson's disease diagnosis from eeg signals. Neural Computing and Applications , 1–7doi:10.1007/s00521-018-3689-5.

Parkinson, J., 1817. An essay on the shaking palsy: London: Whittingham and Rowland for Sherwood. Neely and Jones.

Passos, L.A., Pereira, C.R., Rezende, E.R., Carvalho, T.J., Weber, S.A., Hook, C., Papa, J.P., 2018. Parkinson disease identification using residual networks and optimum-path forest, in: 2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI), IEEE. pp. 000325–000330.

Pereira, C.R., Pereira, D.R., Rosa, G.H., Albuquerque, V.H., Weber, S.A., Hook, C., Papa, J.P., 2018. Handwritten dynamics assessment through convolutional neural networks: An application to parkinson's disease identification. Artificial intelligence in medicine 87, 67–77.

Pereira, C.R., Pereira, D.R., da Silva, F.A., Hook, C., Weber, S.A., Pereira, L.A., Papa, J.P., 2015. A step towards the automated diagnosis of parkinson's disease: Analyzing handwriting movements, in: 2015 IEEE 28th international symposium on computer-based medical systems, IEEE. pp. 171–176.

Phillips, J., Stelmach, G.E., Teasdale, N., 1991. What can indices of handwriting quality tell us about parkinsonian handwriting? Human Movement Science 10, 301–314.

Pujol, J., Junqué, C., Vendrell, P., Grau, J.M., Capdevila, A., 1992. Reduction of the substantia nigra width and motor decline in aging and parkinson's disease. Archives of neurology 49, 1119–1122.

Rana, B., Juneja, A., Saxena, M., Gudwani, S., Kumaran, S.S., Agrawal, R.K., Behari, M., 2015. Regions-of-interest based automated diagnosis of parkinson's disease using t1-weighted mri. Expert Systems with Applications 42, 4506–4516.

Raza, M.A., Chaudry, Q., Zaidi, S.M.T., Khan, M.B., 2017. Clinical decision support system for parkinson's disease and related movement disorders, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1108–1112.

Reeve, A., Simcox, E., Turnbull, D., 2014. Ageing and parkinson's disease: why is advancing age the biggest risk factor? Ageing research reviews 14, 19–30.

Rodriguez, M., Rodriguez-Sabate, C., Morales, I., Sanchez, A., Sabate, M., 2015. Parkinson's disease as a result of aging. Aging cell 14, 293–308.

Rosenblum, S., Engel-Yeger, B., Fogel, Y., 2013a. Age-related changes in executive control and their relationships with activity performance in handwriting. Human movement science 32, 363–376.

Rosenblum, S., Samuel, M., Zlotnik, S., Erikh, I., Schlesinger, I., 2013b. Handwriting as an objective tool for parkinson's disease diagnosis. Journal of neurology 260, 2357–2361.

Ruffini, G., Ibañez, D., Castellano, M., Dunne, S., Soria-Frisch, A., 2016. Eeg-driven rnn classification for prognosis of neurodegeneration in at-risk patients, in: International Conference on Artificial Neural Networks, Springer. pp. 306–313.

Salamon, J., Bello, J.P., 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters 24, 279–283.

Salehinejad, H., Sankar, S., Barfett, J., Colak, E., Valaee, S., 2017. Recent advances in recurrent neural networks. arXiv preprint arXiv:1801.01078 .

Salimans, T., Kingma, D.P., 2016. Weight normalization: A simple reparameterization to accelerate training of deep neural networks, in: Advances in Neural Information Processing Systems, pp. 901–909.

San Luciano, M., Wang, C., Ortega, R.A., Yu, Q., Boschung, S., Soto-Valencia, J., Bressman, S.B., Lipton, R.B., Pullman, S., Saunders-Pullman, R., 2016. Digitized spiral drawing: A possible biomarker for early parkinson's disease. PloS one 11, e0162799.

Saunders-Pullman, R., Derby, C., Stanley, K., Floyd, A., Bressman, S., Lipton, R.B., Deligtisch, A., Severt, L., Yu, Q., Kurtis, M., et al., 2008. Validity of spiral analysis in early parkinson's disease. Movement disorders: official journal of the Movement Disorder Society 23, 531–537.

Smits, E.J., Tolonen, A.J., Cluitmans, L., Van Gils, M., Conway, B.A., Zietsma, R.C., Leenders, K.L., Maurits, N.M., 2014. Standardized handwriting to assess bradykinesia, micrographia and tremor in parkinson's disease. PloS one 9, e97614.

Souza, J., Alves, S., Rebouças, E., Almeida, J., Filho, P.P., 2018. A new approach to diagnose parkinson's disease using a structural cooccurrence matrix for a similarity analysis. Computational Intelligence and Neuroscience 2018, 1–8. doi:10.1155/2018/7613282.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15, 1929–1958.

Stahl, S., 1986. Neuropharmacology of movement disorders: Comparison of drug-induced and spontaneous movement disorders. Movement Disorders .

Stanley, K., Hagenah, J., Brüggemann, N., Reetz, K., Severt, L., Klein, C., Yu, Q., Derby, C., Pullman, S., Saunders-Pullman, R., 2010. Digitized spiral analysis is a promising early motor marker for parkinson disease. Parkinsonism & related disorders 16, 233.

Taleb, C., Khachab, M., Mokbel, C., Likforman-Sulem, L., 2017. Feature selection for an improved parkinson's disease identification based on handwriting, in: 2017 1st

International Workshop on Arabic Script Analysis and Recognition (ASAR), IEEE. pp. 52–56.

Taleb, C., Khachab, M., Mokbel, C., Likforman-Sulem, L., 2018. A reliable method to predict parkinson's disease stage and progression based on handwriting and resampling approaches, in: 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), IEEE. pp. 7–12.

Tanner, C.M., Ottman, R., Goldman, S.M., Ellenberg, J., Chan, P., Mayeux, R., Langston, J.W., 1999. Parkinson disease in twins: an etiologic study. Jama 281, 341–346.

Thomas, M., Lenka, A., Kumar Pal, P., 2017. Handwriting analysis in parkinson's disease: current status and future directions. Movement disorders clinical practice 4, 806–818.

Tisch, U., Schlesinger, I., Ionescu, R., Nassar, M., Axelrod, N., Robertman, D., Tessler, Y., Azar, F., Marmur, A., Aharon-Peretz, J., et al., 2013. Detection of alzheimer's and parkinson's disease from exhaled breath using nanomaterial-based sensors. Nanomedicine 8, 43–56.

Tsanas, A., Little, M.A., McSharry, P.E., Spielman, J., Ramig, L.O., 2012. Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. IEEE Transactions on biomedical engineering 59, 1264–1271.

Vapnik, V., 1998. Statistical Learning Theory. volume 10.

Walton, J., 1997. Handwriting changes due to aging and parkinson's syndrome. Forensic science international 88, 197–214.

Wu, T., Zhang, J., Hallett, M., Feng, T., Hou, Y., Chan, P., 2015. Neural correlates underlying micrographia in parkinson's disease. Brain 139, 144–160.

Yu, N.Y., Van Gemmert, A.W.A., Chang, S.H., 2017. Characterization of graphomotor functions in individuals with parkinson's disease and essential tremor. Behavior Research Methods 49, 913–922. doi:10.3758/s13428-016-0752-y.

Zetusky, W.J., Jankovic, J., Pirozzolo, F.J., 1985. The heterogeneity of parkinson's disease: clinical and prognostic implications. Neurology 35, 522–522.

Zham, P., Kumar, D.K., Dabnichki, P., Poosapadi Arjunan, S., Raghav, S., 2017. Distinguishing different stages of parkinson's disease using composite index of speed and pen-pressure of sketching a spiral. Frontiers in neurology 8, 435.

Zhao, A., Qi, L., Dong, J., Yu, H., 2018. Dual channel lstm based multi-feature extraction in gait for diagnosis of neurodegenerative diseases. Knowledge-Based Systems 145, 91–97.

Zuo, W.L., Wang, Z.Y., Liu, T., Chen, H.L., 2013. Effective detection of parkinson's disease using an adaptive fuzzy k-nearest neighbor approach. Biomedical Signal Processing and Control 8, 364–373.

# A Technical specifications about the databases

## A.1 NewHandPD

The subject H34's age was set at 40 on all tasks except for the SigMea1 where it was set at 50. Therefore, we assumed it was a mistake and set it at 40.

It is actually unclear which database did Pereira et al. (2018) use : at some point they assert that their database comprises only 34 subjects instead of 66 but at the same time they link the database webpage[16] where it is stated that the database is well containing 66 subjects (and one can see for itself after downloading the data). Moreover, Afonso, Pereira, Weber, Hook and Papa (2017), which is almost the same group of authors as Pereira, Pereira, Rosa, Albuquerque, Weber, Hook and Papa (2018), confirms that NewHandPD comprises 66 subjects, although it was published before Pereira et al. (2018). Therefore we assume that Pereira et al. (2018) use the same database as Afonso et al. (2017) and us and that the 34 subjects they talk about were added to the *HandPD* database to form *NewHandPD*.

## A.2 HandPD

We have not analyzed *HandPD* because it is very unclear : on the database webpage[16] and in Passos et al. (2018); Souza et al. (2018); Pereira et al. (2015) it is stated that there is 18 HCs and 72 PDs, although, after downloading the database we can see that the PDs are only numbered from 1 to 38 (and there is no #4). Moreover some PDs have taken up to 8, 12 or 16 spiral tasks instead of 4 like the HCs. The original database might not be available anymore or the subjects' identifier might have been corrupted.

## A.3 PaHaW

The subjects are numbered counting from zero and excluding the subjects that did not perform the spiral task, hereafter.

The tasks spiral, *l*, *lektorka* and sentence from subjects 56, 9, 39 and 67, respectively, started while pen was in air (i.e. not on paper). Although Drotár et al. (2016) state that the recording starts when the pen first touches the paper. Therefore we assume it is a recording error and discarded the in-air points.

In the same way, the subject #43 has a recording problem on its sentence task : the *timestamp* measure jumps to $10^{12}$ on the 12 last points of the recording although the timestamps from all subjects are in the $10^6$ magnitude. Therefore we assume it is a recording error and discarded the 12 last points.

[16] http://wwwp.fc.unesp.br/~papa/pub/datasets/Handpd/

Table 7: Whole-task and strokes' maximum length (in timesteps).

| **task** | **Whole-task** | **Stroke** |
|---|---|---|
| spiral | 16071 | 16071 |
| *l* | 4226 | 752 |
| *le* | 6615 | 1104 |
| *les* | 6827 | 1476 |
| *lektorka* | 7993 | 3568 |
| *porovnat* | 5783 | 2057 |
| *nepopadnout* | 4423 | 2267 |
| sentence | 7676 | 1231 |

Table 7 summarizes the lengths of tasks, whole and split in strokes. These are used for the models described in section 5.1.